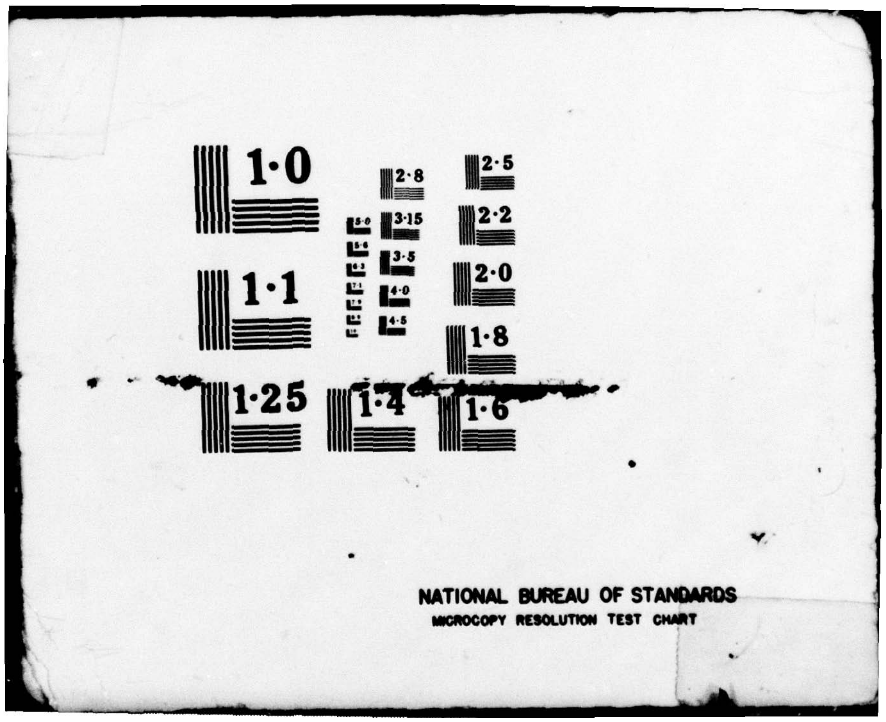
AD-A069 305 MICHIGAN UNIV ANN ARBOR DEPT OF STATISTICS F/G 12/1
STATISTICAL INFERENCE WITH DELAYED DATA.(U)
APR 79 N STARR, M B WOODROOFE DAAG29-76-G-0302
UNCLASSIFIED 014538 ARO-13950.1-M NL

| OF |
AD
A069305

END
DATE
FILMED
7-79
DDC


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5
5·0

ARO 13950.1-M

REPORT DOCUMENTATION PAGE

LEVEL II

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
(6) Statistical Inference with Delayed Data

5. TYPE OF REPORT & PERIOD COVERED
(9) Final rept.

6. PERFORMING ORG. REPORT NUMBER
(14) 014538

7. AUTHOR(s)
Starr, Norman and Woodroofe, Michael B.

8. CONTRACT OR GRANT NUMBER(s)
(15) DAAG29-76-G-0302

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
(11) 30 Apr 79

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
04/30/79

13. NUMBER OF PAGES
10

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
(12) 12p.

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(18) ARO (19) 13950.1-M

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(10) Norman /Starr Michael B. /Woodroofe

18. SUPPLEMENTARY NOTES
The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Optimal stopping; adaptive stopping; search problems, probability of a new species; trap probabilities; one-armed bandit, concomitant data; sequential allocation; adaptive quality control; maximum process; repeated significance tests; level of significance.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The research proceeded in four general areas:

(A) Problems of inference from data accumulating at the random times new species are discovered were studied. Procedures for estimating a variety of characteristics of the population and for predicting the random probability that a next search will uncover a new species were developed. A death process was interposed and estimates of the size of a population (over)

409 719

slt

20. Abstract (cont'd)

and the mean life of its members were obtained.

(B) Results for a one-armed bandit in the presence of concomitant information were developed. These results apply to the sequential allocation of treatments in medical trials.

(C) Optimal and adaptive stopping based on the maximum of a sequence of dependent observations were studied. The results are likely to have application to the choice of components in redundant systems and to adaptive quality control.

(D) Methods have been derived to estimate the true significance level resulting from repeated significance tests.

SUMMARY OF RESULTS

A. A population consisting of distinct species is searched by selecting one member at a time. Each time a new species is discovered we receive an incremental reward, which in statistical applications we regard as a data point to be utilized at the termination of the search in a decision concerning some characteristic of the population. Thus, data is accumulating at the random times we discover new species.

The search may continue indefinitely, but there is a cost $c > 0$ associated with each selection. If after n selections we have discovered $d(n)$ species and elect to terminate the search, our payoff is

$$h(d(n)) - cn ,$$

where h is increasing. Subject to mild conditions on h, it is optimal to terminate the search at the random time

$$\sigma = \text{first } n \geq 0 \text{ such that}$$
$$(h(d(n)+1) - h(d(n))u(n) \leq c$$

where $u(n)$ is the conditional probability we will discover a new species at stage $n+1$ of the search, given the results of the search up to stage n.

As an example of a statistical application of this result, suppose that a specified character is either present in all members of a given species, or else absent. Our objective is to estimate the proportion θ of species carrying the character. If we terminate the search at stage n we must report an estimate $\hat{\theta}_n$ of θ, and incur the normalized loss

$$L(n) = \frac{(\hat{\theta}_n - \theta)^2}{\theta(1-\theta)} + cn ;$$

here c is the cost of each stage of the search. The Bayes strategy with respect to a uniform $(0,1)$ prior on θ is to terminate the search at time

$$s = \text{first } n \geq 0 \text{ such that}$$

$$u(n) \leq cd(n)(d(n) + 1)$$

and to estimate θ with $\hat{\theta}_s$ = the proportion of those species discovered up to time s which carry the character. This strategy is also minimax.

Our results have potential application (for example) to studies of literary vocabulary or codes, species of animals, insects, or microbes in a given area, personality types, and levels of job performance. The obvious limitation in the applicability is that if the composition of the population is unknown, then the random probability $u(n)$ that a new species will be discovered at the next stage of the search is unobservable, and the optimal strategies are unusable. To remedy this deficiency we undertook a detailed study of a class of observable predictors of $u(n)$ based on the frequency of frequencies of distinct species that have been discovered in the search, and of the performance of adaptive strategies which result from "estimating" u(n) with these predictors. We obtained a unique class of linear predictors which were optimal for "estimating" u(n) in a commonly accepted statistical sense and which led to adaptive strategies which appear (from simulations) to perform well against the optimal strategy when the composition of the underlying population is unknown.

In a dissertation (in preparation), Mr. Carlos Lima has extended this study by imposing a death process on the model. Suppose that a population of animals are trapped at given intervals of time in the wild, and that each animal has an associated trap probability (trappability). Of interest are the initial number of animals in the population, the number of species, and their average lifetimes. Standard estimates (maximum likelihood) are biased because of the difficulty of estimating the proportion of the population which is never seen (either because of premature death or small trappability). Predictors of the type we have studied apparently lead to estimators which apparently reduce this bias significantly.

References: [2], [4], and [7]

B. For definiteness, we describe this problem in the language of medical trials; other applications are mentioned later.

We suppose that patients arrive sequentially and may be treated with either a standard treatment (S) or a competing experimental treatment (E). Before deciding which treatment to administer to a particular patient, we observe an associate vector X of relevant concomitant variables; for example, X = (age, sex, severity of disease, time since onset, etc.). Given that $X = x$, let $S(x)$ denote the patient's response if we administer S and $E(x)$ if we administer E. It is assumed that the statistical properties of $S(x)$ (the patient's response to the standard treatment) are known. For the j^{th} patient let

$$Y_j = \begin{cases} S(X_j) & \text{if patient } j \text{ receives } S \\ E(X_j) & \text{if patient } j \text{ receives } E . \end{cases}$$

Our objective is to sequentially administer S and E in such a manner that the average value of

$$\sum_{j=1}^{\infty} \alpha^j Y_j$$

will be a maximum, where $0 < \alpha < 1$. We have adopted a Bayesian approach and developed asymptotically optimal solutions as $\alpha \to 1$. Our results have the following, surprising implications.

1. The approximate solution to the allocation problem is simpler in the presence of the concomitant information X then in its absence;

2. The myopic procedure is asymptotically optimal;

3. The role played by the number N of future patients in earlier work is played by the discount factor α in ours. The exact value of the discount factor α (equivalently, knowledge of N), is of less importance in the presence of concomitant information than in earlier work which did not take account of this information.

Reference: [9]

C. Suppose we may consecutively observe variables $x_1, x_2, \ldots$. For example, each x might represent the quality of a fabricated component or the response to a training session.

(i) Let m_n denote the largest x-value observed after n trials. If sampling is terminated with n trials, and $m_n = y$ say, we receive a payoff $f(n,y)$ which we suppose is increasing in y and decreasing in n. (For example, if there is a cost $c > 0$ for manufacturing a component, and we choose to use the n^{th} component that we manufacture in a system, then $f(n,y) = y - cn$ is an appropriate payoff function). Our objective in this research has been to describe in a form usable in practice the time σ for which the average value of $f(\sigma, m_\sigma)$ will be a maximum. Our results apply to a special form of dependence among the x values (corresponding to an urn sampling without replacement). We prove that

$$\sigma = \text{first } n \geq 1 \text{ such that } m_n \geq \beta_n$$

is optimal (conditions on f are assumed), and provide a simple algorithm for computing the β_n-values. We show also that the myopic strategy

$$\sigma' = \text{first } n \geq 1 \text{ such that } m_n \geq \beta$$

is asymptotically optimal (where β is easily computable).

A student, Dr. Tony Tai, has considered a similar problem in the context of a modified Ehrenfest model (an elastic model). He has derived the optimal stopping time, and also studied in detail the problem of estimating the size of the population from which the model derives via methods of likelihood.

References: [3] and [8]

(ii) Suppose $x_1, x_2, \ldots$ are i.i.d. random variables with absolutely continuous distribution F. Then,by the probability integral transformation $y_i = F(x_i)$, $i=1,2,\ldots$ are i.i.d. Uniform (0,1). Suppose that $F = F_\theta$ is known only up to a vector θ of unknown parameters.

<u>Theorem</u>. If $\hat{\theta}_i = \theta_i(x_1, \ldots, x_i)$ is a complete sufficient statistic and $\hat{\theta}_i^{-1} x_i$ is ancillary (θ^{-1} denotes a group transformation acting on x). Then the sequence

$$\hat{y}_i = F_{\hat{\theta}_i}(x_i) \qquad i=1,2,\ldots$$

are independent with distribution independent of θ (and in general easily derivable).

We expect this result to be useful for constructing new quality control plans (when for example the target value θ is unknown) and for procedures to detect the time at which there may have occurred a shift (in the mean or variablility) of a process.

References: [6]

D. Let $X_1, X_2, \ldots$ denote independent random variables whose densities f_ω, $\omega \varepsilon \Omega$, constitute a natural exponential family and consider testing an hypothesis of the form $H_0\colon \omega \varepsilon \Omega_0$, where Ω_0 is a lower dimensional submanifold of the natural parameter space Ω. If a sample of predetermined size n is taken, then the null distribution of the likelihood ratio statistic

$$\Lambda_n = (\sup_\Omega - \sup_{\Omega_0}) \sum_{i=1}^{n} \log f_\omega(X_i)$$

is approximately that of $\frac{1}{2}\chi_r^2$, where r is the codimension of Ω_0 in Ω. However, if the sample size is allowed to depend on the data in a non-anticipating manner, then the chi-square approximation to the distribution of Λ_n may be poor. In particular, if s is a stopping time the chi-square approximation may badly underestimate the attained significance levels $P\{\Lambda_s > a\}$, where $\omega \varepsilon \Omega_0$ and $a > 0$. If we restrict attention to stopping times which are bounded by a fixed number N, then the latter probability is maximized by the stopping time

$$t = \inf\{n \geq m\colon \Lambda_n > a \text{ or } n \geq N\} ,$$

where m is an initial sample size (so chosen that Λ_n is finite for all $n \geq m$).

We have obtained approximations to $P\{\Lambda_t > a\}$ for $\omega \varepsilon \Omega_0$ and large a in a context which is sufficiently general to include many multivariate applications. Our results show that the actual attained significance level may exceed the nominal $\Pr\{\chi_r^2 > s\}$ by a factor of 10-20 times for $50 \leq N \leq 300$ when r is small. These results confirm and extend the numerical work of Armitage et al. and Siegmund's treatment of repeated t-tests.

References: [1], [5], [10], [11] and [12]

REFERENCES

[1] Armitage, P., C.K. McPherson, and B.C. Rowe (1969), "Repeated significance tests on accumulating data," J. R. Statist. Soc., A 132, 235-244.

[2] Bartold, S. and Starr, N. (1979), "Optimal and adaptive stopping times in the search for new species," in publication to JASA.

[3] Chen, Wen-Chen and Starr, N. (1979), "Optimal stopping in an urn," in publication to Ann. of Prob.

[4] Lima, Carlos (1979), "Inference from trapping data," Ph.D. dissertation, Department of Statistics, University of Michigan, to be submitted for defense May 1979.

[5] Siegmund, D. (1977), "Repeated significance tests for a normal mean, Biometrika, Vol. 64, 177-190.

[6] Starr, N. "Adaptive quality control," in progress.

[7] Starr, N. (1979), "Linear estimates of the probability of discovering a new species," in publication: Ann. Math. Statistics.

[8] Tai, Tony (1978), "Optimal and adaptive stopping for a generalized urn model," Ph.D. dissertation, Department of Statistics, University of Michigan.

[9] Woodroofe, M.B. (1979), "A one-arm bandit with a concomitant variable," submitted for publication in JASA.

[10] Woodroofe, M.B. (1976), "Frequentist properties of Bayesian sequential tests," Biometrika, Vol. 63, 101-110.

[11] Woodroofe, M.B. (1978), "Large deviations of likelihood ratio statistics with applications to sequential testing," Ann. Statist. Vol. 6, 72-84.

[12] Woodroofe, M.B. (1978), "Repeated significance tests," Technical Reprot #83, Department of Statistics, University of Michigan.